Synergistic Discovery and Design

Jennifer Roberts, Ph.D.
I2O Program Manager

Proposer's Day

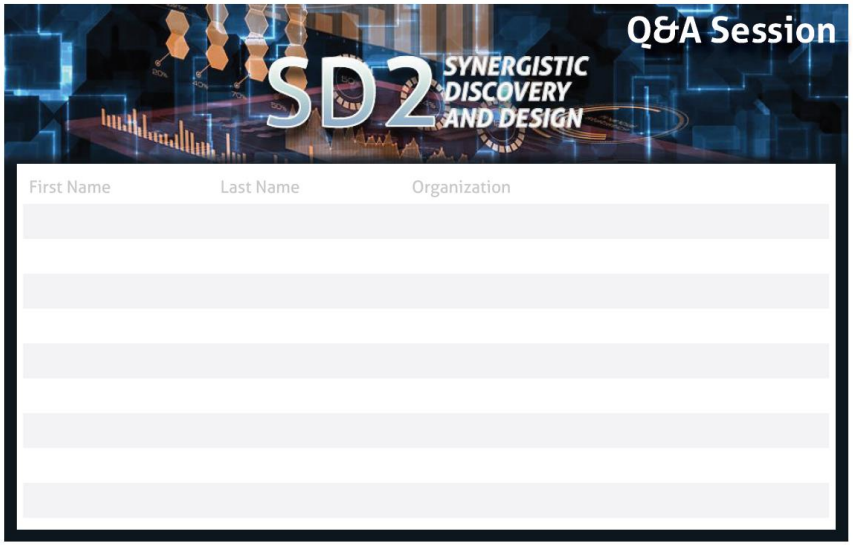
November 10, 2016

Outline

- Introduction/Background
- Synergistic Discovery and Design Approach
- Challenge Problem Overview
  - Notional Challenge Problem Examples
- Program Description
  - TA1: Data Centric Scientific Discovery
  - TA2: Design in the Context of Uncertainty
  - TA3: Hypothesis and Design Evaluation
  - TA4: Data and Analysis Hub
  - TA5: Challenge Problem Integrator
  - Quarterly Challenge Problem Cycle
- Collaborative Program Structure
- Program Wide Milestones
- SD2 Program Logistics
- SD2 Team

Questions?

- Fill out a question card



- Send an email to: SD2@darpa.mil
- Attendees may submit questions via index cards until 10:00 AM
- DARPA Q&A responses: 1:20PM

Synergistic Discovery and Design (SD2)

Develop data-driven methods to accelerate design
in domains that lack complete models.

Model discovery will enable rapid refinement of designs in domains with the following characteristics:

- Millions of unpredictable, interacting components for which we lack robust models
- Partially observable internal states
- Unexpected design failures due to small perturbations
- Operational envelopes with an unknown number of engineering variables and an underdetermined degree of interaction between variables

Domains of interest include synthetic biology, neuro-computation, and polymer chemistry

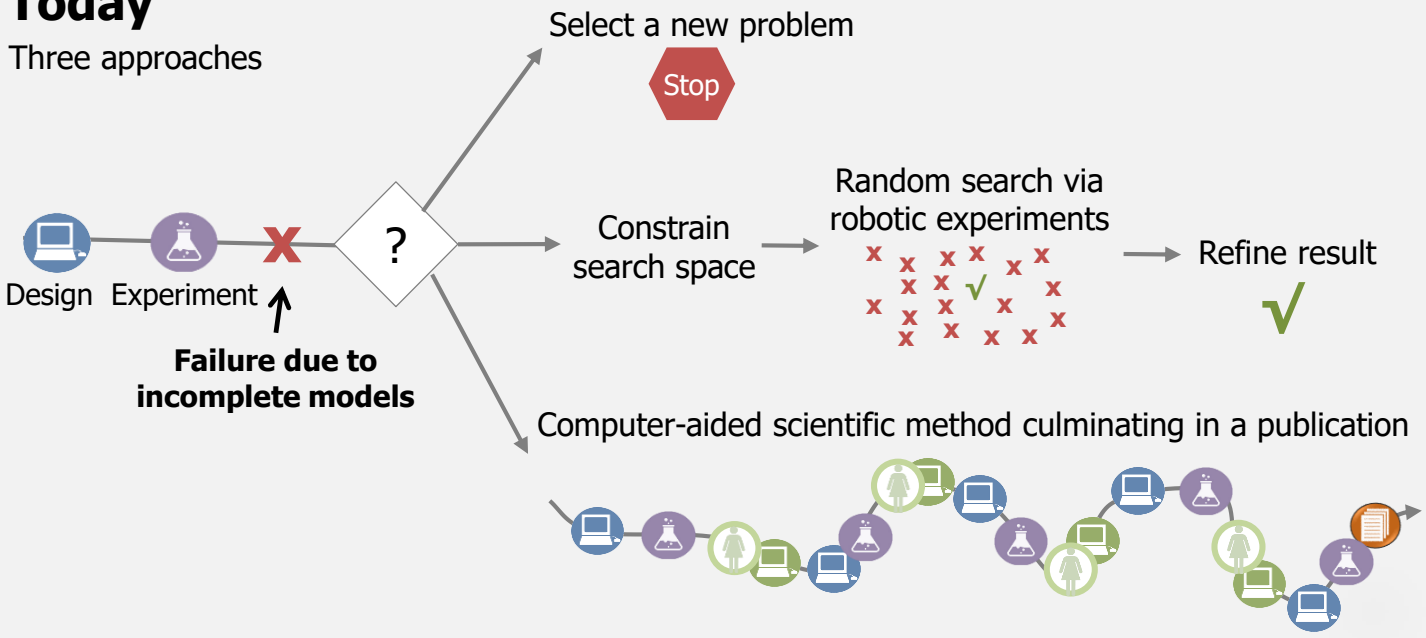
Domains not of interest include those with high-fidelity simulations, such as aeronautics, automobiles, and integrated circuits

Approaches to design in domains with incomplete models

Notional Design Challenge: Create a biological circuit to absorb nuclear waste

Today

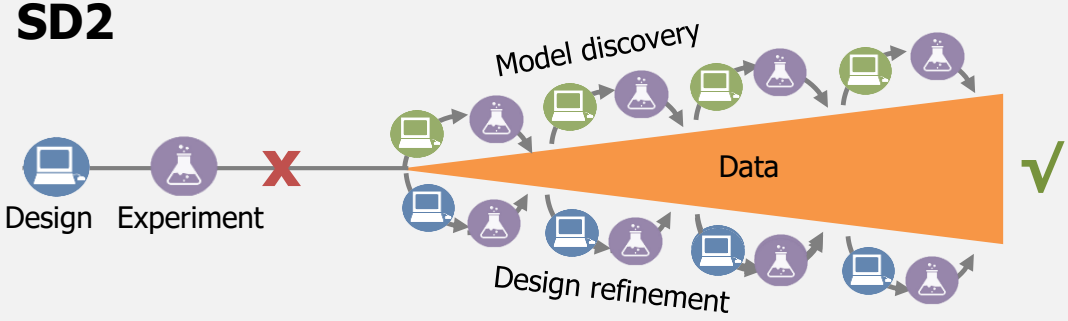
Three approaches



Outstanding questions

- Why did the design fail?
- What caused the success?
- Are the results reproducible?
- Could the process be more efficient?

SD2

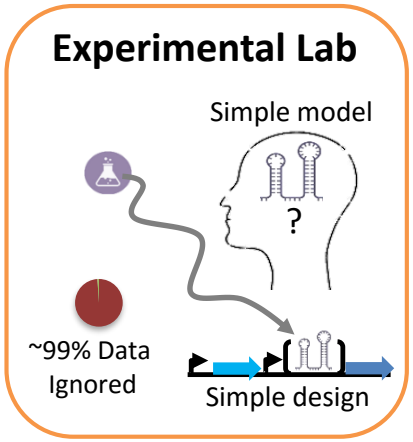


Outcomes

- I know why designs fail
- I know what causes success
- I am confident results are reproducible
- I efficiently converged on successful designs

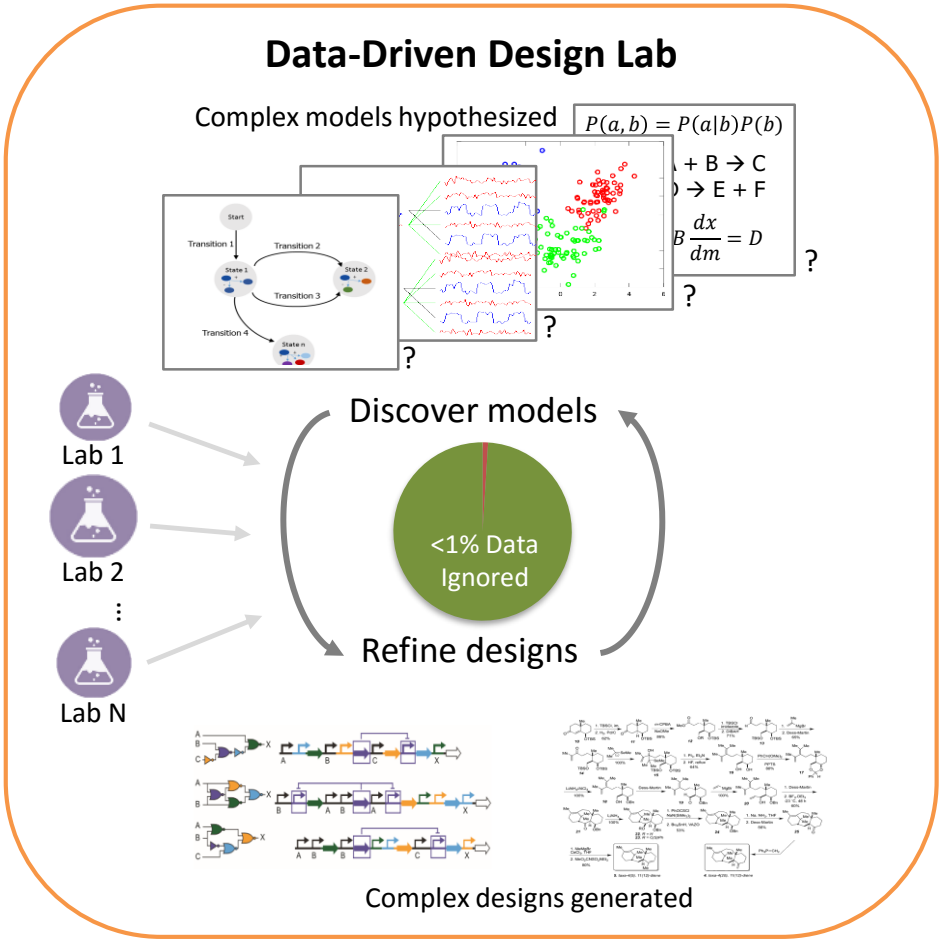
SD2 will increase model and design complexity by enabling data-driven design and discovery at scale

Today



- Analyze GB of data from a single lab
- Examine 10s of engineering variables per design
- Discover one publication-worthy model refinement every few years

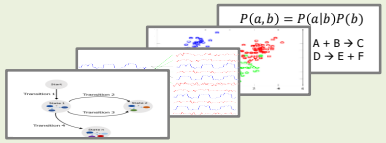
SD2



- Analyze 100s TB of data from multiple labs
- Examine 1x10^{6}s engineering variables per design
- Discover multiple publication-worthy model refinements per year

Synergistic Discovery and Design approach

TA1: Data-Centric Scientific Discovery

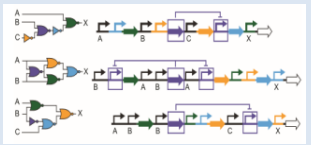


Extract scientific knowledge and theory directly from experimental data at scale

Experimental outcomes

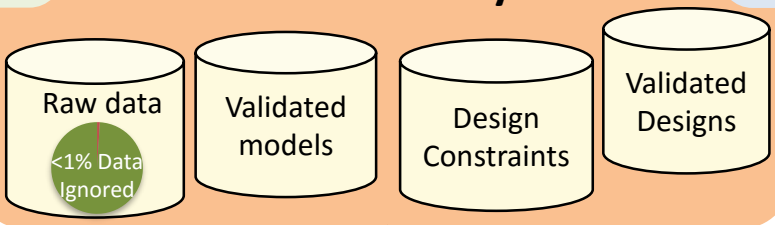
Validated models

TA2: Design in the Context of Uncertainty



Automate development of robust designs despite incomplete scientific knowledge

TA4: Data and Analysis Hub



Scientific hypotheses

Observed behavior (Successes and failures)

Design/ Experimental plan

TA3: Hypothesis and Design Evaluation

Virtualize experimental workflows to facilitate reproducibility

TA5: Challenge Problem Integrator

Direct generation of quarterly challenges problems

Notional challenge problems

Synthetic Biology: Design 30 component circuit for nuclear waste absorption

Polymer Chemistry : Design a polymer that constricts in response to toxin

Challenge problem overview

Program wide evaluations of SD2 methods will involve a series of real world design challenges that today's scientific community does not yet know how to solve.

Intent: Develop data-driven methods that can generate new scientific discoveries and design new capabilities that are not currently found in literature and are outside the limits of current science.

Criteria for Success

- Solve the posed challenge problems at a rate faster than current publication cycles
- Decrease the expected time to convert an academic finding into a robust, industrial design or process

Notional challenge problems for alternative domains

Neurochemistry: Design chemical structures that penetrate the blood-brain barrier

***Polymer Chemistry:* Design a polymer that constricts in response to toxin**

Medicine: Design compounds with efficacy against emerging diseases

Social Behavior: Discover factors that cause a person to be vulnerable to radicalization

Epidemiology: Discover travel patterns that promote the spread of Zika virus

Genetics: Discover genetic factors that lead to rare diseases

Medical Records: Discover factors contributing to disease onset

Nuclear Physics: Discover conditions that enhance the speed of radioactive decay

Neuroscience: Discover neural pathways associated with cognitive processes

Meteorology: Reduce error of hurricane trajectory predictions

Sociology: Discover causes of growth stunting due to severe malnutrition

Cancer Biology: Discover similarities in signaling behavior within 1000 cancer cell lines

Education and Training: Design training personalization regimes that improve retention

Notional challenge problems

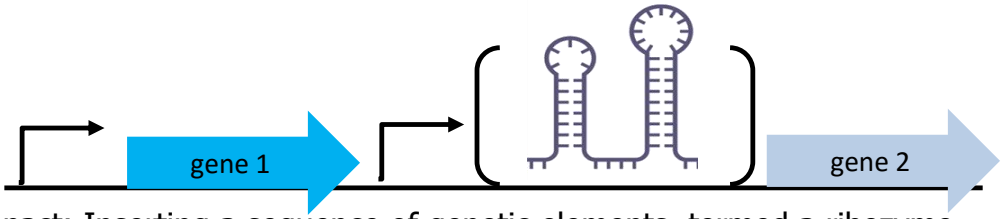
Challenge Problem Guidelines

- Facilitate robust design at scale
- *Do not present IP or data sharing restrictions*
- Enable exploration of topics relevant to DoD interests such as reproducibility and robustness
- Lead to discovery and design hypotheses that can be experimentally tested in days, weeks, or months

	Synthetic Biology (High Throughput Application Domain)	**Material Science** (Low-Mid Throughput Application Domain)	**Criteria for Success**
Phase 1	Reproduce published designs in simple organisms	Synthesize polymers from different monomer chemistries	Reduce variability across labs due to experimental conditions
Phase 2	Adapt published designs to similar simple organisms	Generate polymers with predicted performance for a specified application	Discover root causes for experimental surprises due to flawed scientific models
Phase 3	Discover novel designs for more complex organisms	Discover novel polymers for DoD relevant applications	Discover non-canonical mechanisms and use them for design

Illustrative example of data-centric scientific discovery

Representative Discovery: Biological insulators increased biological circuit design success rates from 5% to 99% (Lou et al., Nature Biotech, 2012).



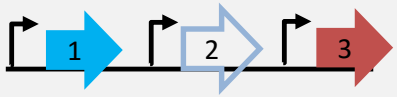
Impact: Inserting a sequence of genetic elements, termed a ribozyme-based biological insulator, enables expression of any two adjacent genes

2 years

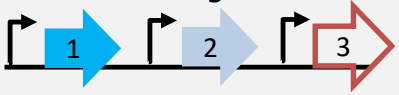
Data-driven techniques would have accelerated the discovery of biological insulators

1. Discover categories

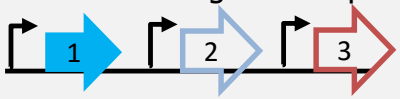
Behavior one: genes 1 & 3 expressed



Behavior two: genes 1 & 2 expressed

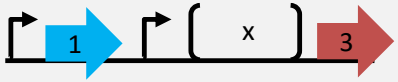


Behavior three: gene 1 expressed

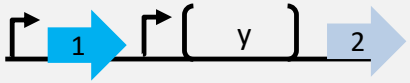


2. Identify qualitative relationships

Gene 1 & 3 expressed when separated by sequence x

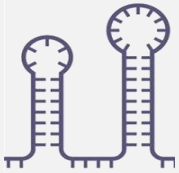


Gene 1 & 2 expressed when separated by sequence y



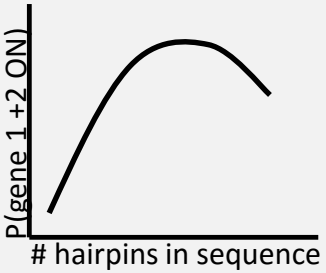
3. Find causes of desired behavior

Sequence x & y must contain at least N hairpin structures



Sample sequence

4. Quantify



3 months

Hairpin image: http://mmbr.asm.org/content/74/4/570/F4.expansion.html

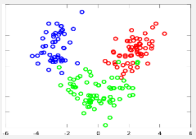
TA1: Data-Centric Scientific Discovery

Develop computational workflows and algorithms to automatically detect experimental surprise, extract patterns from experimental data at scale, and use patterns to refine models

Research Challenges

1. Discover categories of engineering variables
2. Identify qualitative relationships
3. Find stable regions of operation
4. Quantify relationships between engineering variables

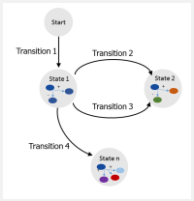
Similar traits



Transformations

A + B → C
C + D → E + F

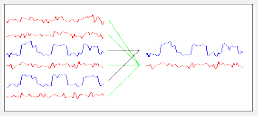
States and transitions



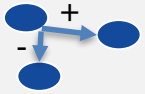
Differential equations

$$A\frac{d^2x}{dm} + B\frac{dx}{dm} = D$$

Similar behavior



Influence models



Rules

If {X has a negative effect on Y}, then {X inhibits Y}

Probability functions

$$P(a,b) = P(a|b)P(b)$$

Anticipated Approaches:

- Combined workflows
  - Clustering + cross correlation + Markov models + Baum-Welch parameter estimation
- Novel methods within a single mathematical formalism
  - Hierarchical generative model for categories, relationships, conditions, and quantifications
- Algorithmic approximations that scale approaches to petabytes of data

TA1 guidance

Goal: Develop computational methods that convert experimental data into scientific models. Scientific models should lead to testable experimental hypotheses and provide information that design algorithms can use for computation.

Technical Interests:

- Novel computational methods that address the subtasks described on the previous slide
- Reduced human intervention
- Methods to increase model complexity at a rate commensurate with empirical evidence
- Discovery at scale

Special Notes: Teams will be expected to demonstrate methods on government-provided data and data provided by TA3. Competitive proposals will provide examples of notional discoveries.

Team Characteristics:

- Expertise: Computer science, mathematics, data science; Application domain expertise optional
- Multiple teams expected
- Team size ~ $800 - $1.5M per year

Interactions with other TAs

- TA2: TA1 provides design constraints, models, and levels of uncertainty to TA2
- TA3: TA1 tests and demonstrates methods on data generated by TA3
- TA4: TA1 uses TA4 infrastructure to demonstrate methods
- TA5: TA1 provides insights extracted from the data to propose challenge problems

TA2: Design in the Context of Uncertainty

Identify and extend engineering and planning formalisms that enable automated refinement of flawed, yet well-principled designs

TA5: Design challenge problem

Synthetic Biology ex.: Design 30 component circuit for nuclear waste absorption
Polymer Chemistry ex.: Design a polymer that constricts in response to toxin

TA4: Hub

Relevant empirical models?

Yes → Validated models

No → Models with uncertainty

Experimental surprise?

Yes → New Data

No

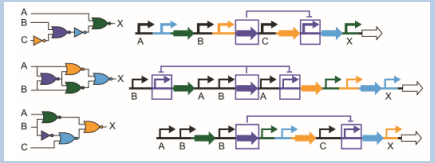
TA1: Discovery

Discover system patterns, outliers and model refinements from experimental data

Refined Models

TA2: Design

TA2.A: Design despite incomplete knowledge by dynamically incorporating new scientific discoveries



TA2.B: Propose experiments with maximal information value by using planning algorithms

$$\underset{x_i \in \text{experiments}}{\arg\max} \left\{ \sum_{i=1}^{N} \mathrm{E}[information_{x_i}] \right\}$$

New Data (Experimental successes and failures)

TA3:Empirical evaluation

Collect new data

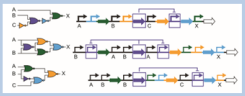
Robust Design

TA2 guidance

Goal: Develop algorithms that provide novel design capabilities in domains with incomplete models.

TA2: Design

TA2.A: Design despite incomplete knowledge by dynamically incorporating new scientific discoveries



TA2.B: Propose experiments with maximal information value by using planning algorithms

$$\underset{x_i \in \text{experiments}}{\arg\max} \left\{ \sum_{i=1}^{N} \mathrm{E}[information_{x_i}] \right\}$$

Technical Interests:

- Algorithms that design novel capabilities within a single application domain
- Experimental planning algorithms that optimally use experimental resources to test proposed designs
- Production of computer-readable experimental plans for TA3
- Algorithms that incorporate TA1 models

Special Notes: Teams will be expected to demonstrate methods on government-provided data and data generated by TA3. Competitive proposals will provide examples of notional designs and experimental plans. Proposals for design within alternative application areas should identify TA3 labs that could generate relevant data.

Team Characteristics:

- Expertise: Computer science, planning algorithms, background in empirical design in experimental domains (e.g. biology, chemistry)
- Multiple teams expected
- Team size ~ $800 - $1.5M per year

Interactions with other TAs

- TA1: TA2 incorporates information from TA1 models
- TA3: TA2 uses TA3 APIs to submit experimental plans and uses TA3 data to evaluate methods
- TA4: TA2 uses TA4 infrastructure to demonstrate methods
- TA5: TA2 provides insights extracted from the data to propose challenge problems

Joint TA1/TA2 guidance

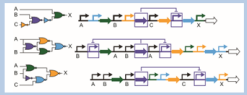
Goal: Develop algorithms that provide novel discovery and design capabilities in domains with incomplete models.

Technical Interests:

- Novel computational methods that address the TA1 subtasks
- Experimental planning algorithms that optimally use experimental resources to validate TA1 models
- Optional: TA2.A algorithms that design novel capabilities within a single application domain
- Production of computer-readable experimental plans for TA3

TA2: Design

TA2.A: Design despite incomplete knowledge by dynamically incorporating new scientific discoveries



TA1: Discovery

Discover system patterns, outliers and model refinements from experimental data

TA2.B: Propose experiments with maximal information value by using planning algorithms

$$\underset{x_i \in \text{experiments}}{\arg\max} \left\{ \sum_{i=1}^{N} \mathrm{E}[information_{x_i}] \right\}$$

Required for TA1/TA2 Joint Proposals

Special Notes: Teams will be expected to demonstrate methods on government-provided data and data generated by TA3. Competitive proposals will provide examples of notional discoveries, designs, and experimental plans.

Team Characteristics:

- Expertise: Computer science, mathematics, data science, planning algorithms; Background in experimental domains (e.g. biology, chemistry) is recommended
- Multiple teams expected
- Cost per TA should be separated and commensurate with the proposed technical approach

Interactions with other TAs

- TA1: TA2 incorporates information from TA1 models
- TA3: TA2 provides experimental plans to TA3 and uses TA3 data to test and demonstrate methods
- TA4: TA2 uses TA4 infrastructure to demonstrate methods
- TA5: TA2 provides insights extracted from the data to propose challenge problems

TA3: Hypothesis and Design Evaluation

Virtualize experimental workflows to facilitate access to high-throughput experimental hardware and enable comparison of data across geospatially distributed laboratories

Today

Manual Workflows

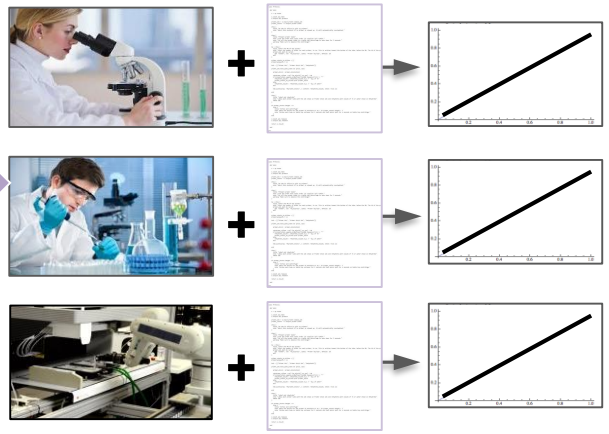
Different Users

Different Results

SD2

Experimental Workflow Code

With Computer Readable Workflows



Different Users

Correlated Results!

Outcome

High Quality Data

to

Discover models

Refine designs

Images from: http://www.uwbiofab.org/

TA3 guidance

Goal: Experimentally evaluate the designs and hypothesized system knowledge generated by TA1 and TA2 to test for reproducibility and robustness

Biology Application Domain (High Throughput)	Alternative Application Domains (Low-Mid Throughput)
Technical Interests: • Labs in 3 geographically distinct areas • Equipment and staff to perform experiments with different types of single-celled organisms and cell lines • High-throughput robotics equipment: 100 GB to 1 TB data per day • Breadth of experimental capabilities • Sensor instrumentation • Computer readable workflow language	**Technical Interests:** • Instrumentation for data breadth: MBs to 100s GBs data per day • Sensor instrumentation • Computer readable protocol language • Novel, high-impact data collect
Team Characteristics: • 1 team or 2-3 primes with a teaming agreement • Team size <$3M base per year (total across all high-throughput teams) • Include options for additional data	**Team Characteristics:** • 1 team or 1-2 primes with a teaming agreement • Team size <$500k base per year (total across all low-mid throughput labs in the same domain) • Include options for additional data

Special Notes: Proposals should describe design challenges that could be addressed using the data

Interactions with other TAs

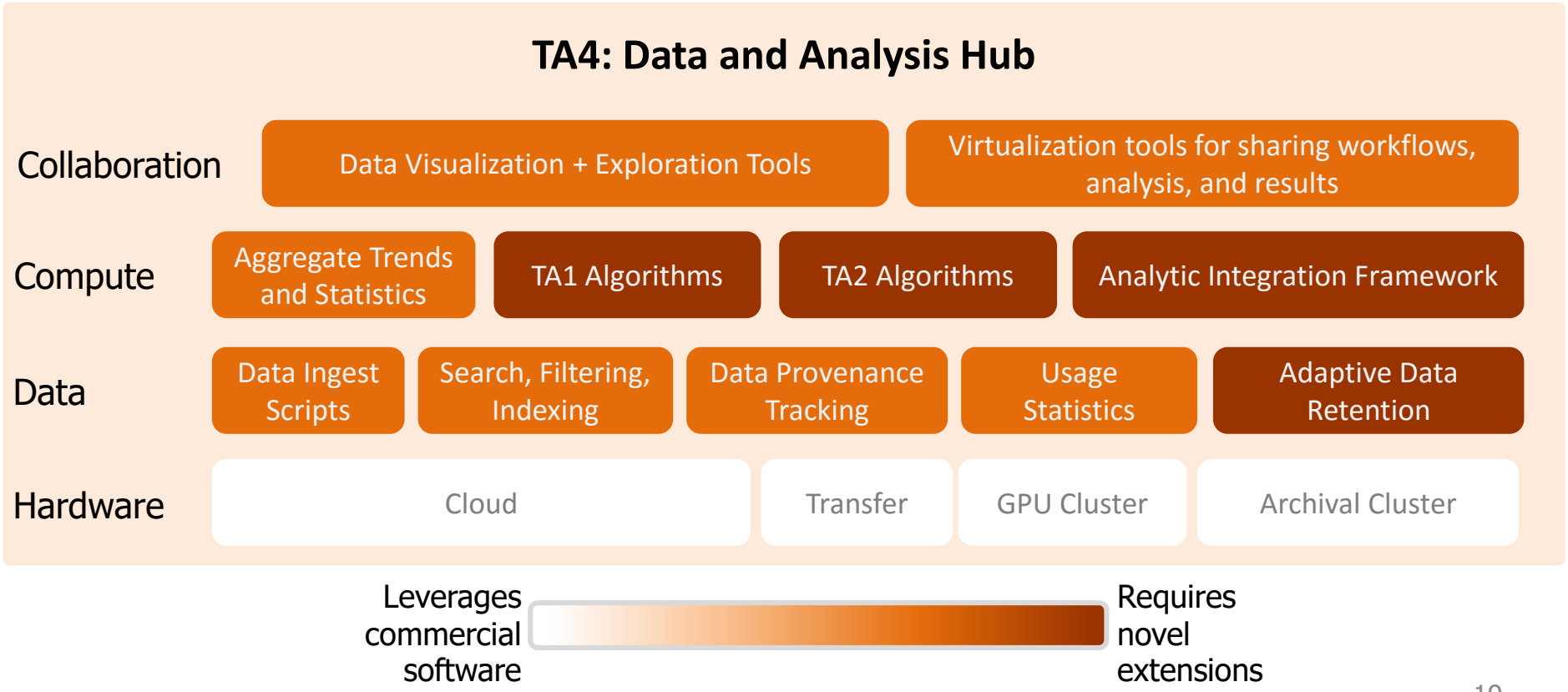
TA1 & TA2: TA3 provides experimental data for method/algorithm development
TA2: TA3 receives experimental plan from TA2
TA4: TA3 provides experimental data to TA4
TA5: TA3 provides subject matter expertise to propose challenge problems

TA4: Data and Analysis Hub

Extend open source tools to virtualize access to physical resources, data, and results in order to link multiple research communities

Research challenges:

- Develop adaptive data management tools that adjust data access speed based on usage
- Develop a flexible TA2/TA3 integration framework that supports experimentation
- Extend collaborative virtualization tools to enable immediate sharing of analytic environments
- Analyze usage statistics to evaluate which collaboration tools provide the most value



TA4 guidance

Goal: Create computational infrastructure that provides alternatives to the way people do science today

Technical Requests:

- Proposals should be organized into the following sections:
  - Hardware plan
  - User-management plan
  - Data management
  - Analytic environments
  - Integration
  - Collaboration tools
  - Maintenance
- Clearly distinguish between existing commercial or open source tools and novel extensions
- Novel ways to support collaborative science are highly encouraged

Team Characteristics:

- Expertise: Cloud and GPU infrastructure management, data management, ETL of complex data, algorithm integration, and collaboration tools
- 1 team or 2-3 teams with a teaming agreement expected
- Team size $3M per year base (total across all teams) with options for additional storage and compute

Interactions with other TAs

- All TAs: TA4 to provide storage and compute as well as tutorials on infrastructure, data processing techniques, and collaboration tools
- TA1 and TA2: TA4 to provide APIs to run methods during data ingest
- TA3: TA4 to provide hardware for TA3 data transfers; TA4 will work with TA3 to establish data formats and facilitate timely ingest of experimental data

TA5: Challenge Problem Integrator

Coordinate generation of quarterly challenges problems for synergistic model discovery and design refinement at scale

Roles and responsibilities:

- Elicit knowledge from domain experts from TA1-TA4 to establish quarterly challenge problems that drive the development of tools for data-driven design in domains that lack models.
- Organize technical agenda for quarterly working PI meetings to coordinate experimental data sharing, collaboratively learn scientific discoveries, and revise program challenge problems
- Act as an evaluator to assess whether SD2 methods support discovery and design beyond what is otherwise possible today
- Evaluate TA1 and TA2 algorithms to human-data analysis capabilities in order to determine which subtasks humans perform best, what subtasks algorithms perform best, and how the overall effect might be amplified via novel human-algorithm partnerships

TA5 guidance

Goal: Work with domain experts throughout the program to establish quarterly challenge problems to drive development of automated data-driven methods

Technical Interests:

- Notional process for challenge problem selection
- Collaborative approach backed by scientific research
- Notional plan for a working PI meeting
- Notional metrics for program evaluation

Team Characteristics:

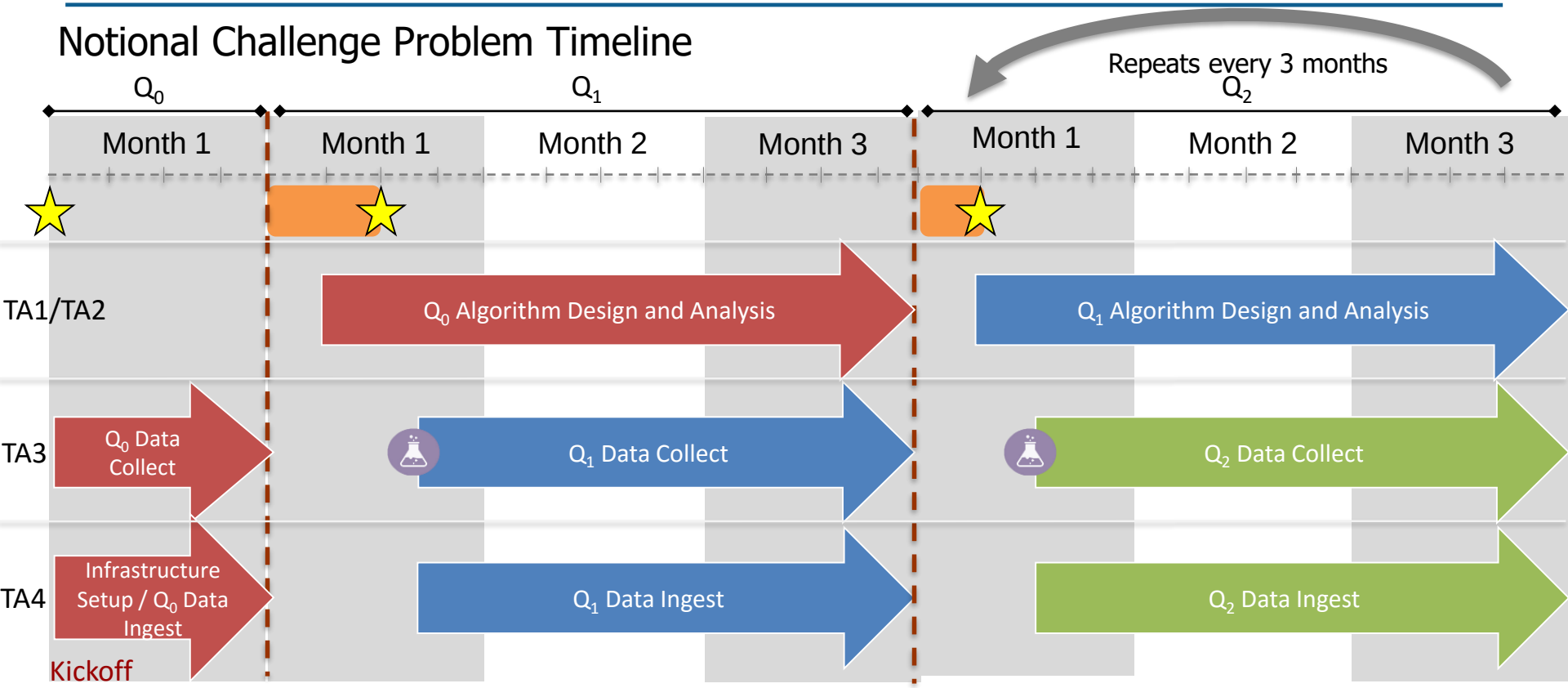
- Expertise: Knowledge elicitation, evaluation of processes that involve both humans and machines
- 1 small team expected

Interactions with other TAs

- TA5 will facilitate collaboration among all SD2 performers
- TA1: TA5 will deliver discovery challenges to TA1
- TA2: TA5 will deliver design challenges to TA2
- TA3: TA5 will prioritize experiments proposed by TA1 and TA2

Quarterly challenge problem cycle

Notional Challenge Problem Timeline



TA5 will help SD2 to achieve working PI meeting objectives:

- Facilitate collaboration among performers
- Coordinate data management efforts
- Exchange insights gained from validation experiments
- Evaluate the effectiveness of discovery and design algorithms
- Refine challenge problems
- Prioritize experiments for the next quarter

Working PI meeting

Challenge problem decision

Experimental plan delivered

Program wide milestones

- Analyze 100s TB to PB of data from multiple labs
- Examine $1x10^6$s engineering variables per design
- Discover multiple publication-worthy model refinements per year
- Increase automation and reduce human intervention to refocus human efforts on the science
- Accelerate the rate of discovery and design with novel computational methods

Level of effort summary by task

For each task and subtask, identify the skill set(s) of the individuals who will perform the task.

SOW Task		Duration (months)	Intensity (hrs/mo)	Labor Hours for Prime							Labo	
				Sr	Skill set(s)	Mid	Skill set(s)	Jr	Skill set(s)	Total	SubC-Sr	Skill set(s)
1.1.0	<Phase 1 Task 1 name>	**7**	**135**	**240**		**680**		**24**		**944**	**-**	
1.1.1	<Subtask 1.1.1 name>	4	90	80		280		-		360	-	
1.1.2	<Subtask 1.1.2 name>	3	195	160		400		24		584	-	
1.2.0	<Phase 1 Task 2 name>	**6**	**385**	**108**		**400**		**1,800**		**2,308**	**1,400**	
1.2.1	<Subtask 1.2.1 name>	3	656	48		320		1,600		1,968	600	
1.2.2	<Subtask 1.2.2 name>	3	113	60		80		200		340	800	
:	:	:	:	:		:		:		:	:	
		Phase 1 Total Hours		**348**		**1,080**		**1,824**		**3,252**	**1,400**	
Phase 1 Costs *First column is prime, second is total subcontractor, third is total consultant, fourth is total*								**Travel**		*$ 44,000*	*$ 12,000*	
								Materials & Equipment		*$ 8,000*	*$ -*	
2.1.0	<Phase 2 Task 1 name>	**8**	**100**	**176**		**560**		**64**		**800**	**100**	
2.1.1	<Subtask 2.1.1 name>	7	51	96		240		24		360	100	
2.1.2	<Subtask 2.1.2 name>	4	110	80		320		40		440	-	
2.2.0	<Phase 2 Task 2 name>	**6**	**417**	**180**		**520**		**1,800**		**2,500**	**1,240**	
2.2.1	<Subtask 2.2.1 name>	4	435	140		400		1,200		1,740	400	
2.2.2	<Subtask 2.2.2 name>	4	190	40		120		600		760	840	
:	:	:	:	:		:		:		:	:	
		Phase 2 Total Hours		**356**		**1,080**		**1,864**		**3,300**	**1,340**	

SD2 program logistics

TIPS

- Read the anticipated BAA carefully
- Ask questions and check the SD2 FAQ document
  - Email questions to: SD2@darpa.mil
  - FAQ: http://www.darpa.mil/work-with-us/opportunities

  Formation of teams
  - Teaming is strongly encouraged for TA3 and TA4
  - Teaming website is available at: https://www.schafertmd.com/darpa/i2o/sd2/teaming
  - There is no bias for teams internal to one institution or across multiple institutions
    - But, effective communication and collaboration between teams is necessary

The SD2 Team

Jennifer Roberts, Ph.D.
Program Manager

Samuel Aldrich, Diana Chung
Business & Financial

Melissa St. Amand, Ph.D.

Jenica Patterson, Ph.D.

www.darpa.mil